科技部科技基础性工作专项资金重大项目
研究成果

项目名称：我国数字图书馆标准规范建设

子项目名称：数据检索与应用标准规范研究

项目编号：2002DEA0018

研究成果类型：研究报告

成果名称：OAI-PMH 项目案例分析——Kepler 项目

成果编号：CDLS-S07-003

成果版本：总项目组推荐稿

成果提交日期：2003 年 8 月

撰写人：牛振东（国家图书馆）

朱先忠（国家图书馆）

项目版权声明

作者声明

本报告作者谨保证本作品中出现的文字、图片、声音、剪辑和文后参考文献等内容的真实性和可靠性，愿按照《中华人民共和国著作权法》，承担本作品发布过程中的责任和义务。科技部有关管理机构对于本作品内容所引发的版权、署名权的异议、纠纷不承担任何责任。

《我国数字图书馆标准规范建设》课题组网站（http://cdls.nstl.gov.cn）作为本报告的第一发表单位，并可向其他媒体推荐此作品。在不发生重复授权的前提下，报告撰写人保留将经过修改的项目成果向正式学术媒体直接投稿的权利。

OAI-PMH 项目案例分析——Kepler 项目

目　录

1. 项目简介

Kepler[1] – A Digital Library For Building Communities。Kepler的目标是使任何一个用户都可以很容易地通过“archivelet”（一个自包含、自安装的软件系统，该系统的功能是实现OAI[2]协议中的数据提供者）对自己的文档进行发布[3]。Kepler的archivelet安装、使用和维护都非常简单。使用Kepler的工具，可以构建与OAI-PMH兼容的数据提供者系统，而不需要复杂的OAI-PMH软件的安装，其特点如下：

- 可配置的团体标准 Configurable community standards；
- 定制的发布和搜索服务 Tailored publication and search services；
- 广泛地和快速地分发 Broad and fast dissemination；
- 可以和其他团体进行交互 Interoperable with other communities；

Kepler 最初的概念是基于并从 OAI 演化而来。OAI 可以在数字图书馆之间建立互联关系，OAI 框架分为数据提供者和服务提供者，而数字图书馆属于数据提供者，并同意将其拥有的元数据资源以某种标准进行发布，使服务提供者基于 OAI-PMH 协议去收集有数据提供者提供的元数据信息。基于 OAI 的 Kepler 框架支持所谓的“个人数据提供者”或“archivelets”，Keple 框架的目标是：

- 可以满足在一般大学中普通的研究者及时地向广大的读者发布其科研成果及论文；
- 可以使普通大众无缝的访问所有发布的资料。

2. Kepler 框架－Kepler Framework

Kepler 根据数字图书馆建设的特点，即使用和控制的便捷，将研发重点放在了发布工具（publication tools）上，从而创建了 archivelet，是独立于平台的，可以安装在工作站或 PC 机上的软件包，而不是只能在个别组织系统内部安装的软件系统，如由 eprints.org 开发的，与 OAI 兼容的软件包。Archivelet 具有非常方便的图形用户界面，可以制作与 OAI 兼容的数据提供者系统。Archivelet 的设计初衷是尽量避免过分依赖于其他软件系统，因此 archivelet 使用本地的文件系统存储少量的对象数据而不使用商业数据库系统。对 archivelet 的支持中，注册服务所起的作用要比常规的 OAI 中的注册服务器起的作用更大。OAI 的注册服务对与 OAI 兼容的文档库和当前注册处理都是手动的，不同于组织级的数据提供者，archivelet 将在活动与非活动状态间具有更灵活的转换方式。Kepler 借鉴了 Napster 和及时消息（instant-messenger）模式的中心服务器跟踪活动客户段的方式。

OAI 从技术层面和组织层面描述了元数据获取框架，这个框架设计用于对分

布的文档库中存储的内容进行发现。OAI 技术框架由两个部分组成，①一套元数据元素的定义（如 OAI 使用的都柏林内核）；②定义用于提取文档元数据的公用协议。同时 OAI 还定义了两个不同的参与者：数据提供者和服务提供者。目前 OAI 框架只支持组织级的数据提供者。基于 OAI 的 Kepler 框架所支持的 archivelet 是面向众多小的出版者的。Kepler 框架提高了个人出版者的技术文档的分发速度。图 1 列出了 Kepler 框架的组成：OAI 兼容的仓储、出版工具（publishing tool）、注册服务、服务提供者。

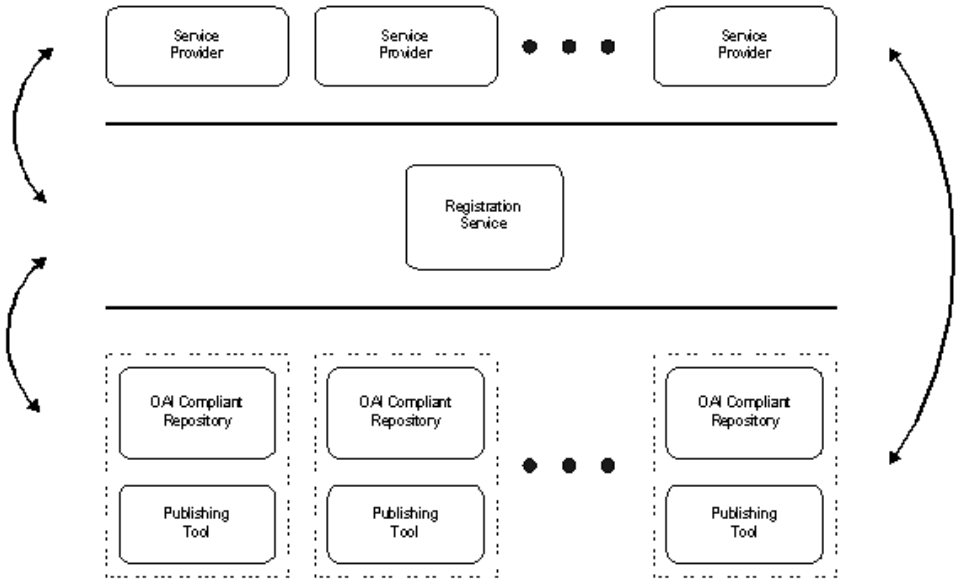

图 1　Kepler 框架

从图中可以看到，OAI 兼容的仓储是和出版工具在一起的，也就是 Kepler 框架的 archivelet。注册服务是用于跟踪注册的 archivelet 的可获得的状态信息。服务提供者提供更高级别的服务，如允许读者从所有注册的 archivelet 中搜索已出版的文档资料。

Kepler 框架支持两种类型的使用者，一是使用 archivelet 出版工具的个人出版者；二是对搜索已出版文档感兴趣的一般使用者。前者只与出版工具交互，后者是通过通用浏览器与 OAI 兼容的服务提供者交互。Kepler 框架类似于基于点对点（Peer-to-Peer）网络模式的代理程序，典型的，一个用户既可以是数据提供者，也可以是使用服务提供者所提供的服务的使用者。如图 2 所示。

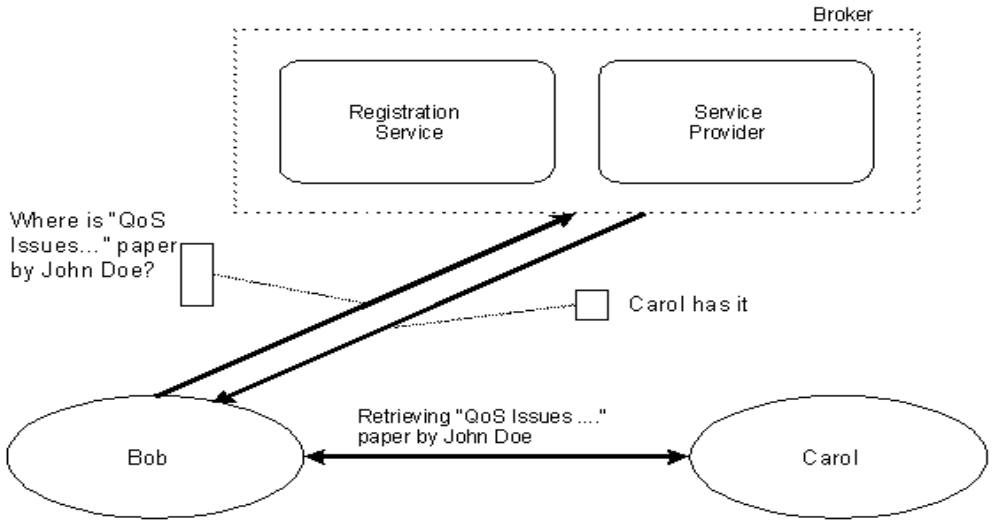

图 2　Kepler 框架和 Peer-to-Peer 网络模式

目前 Kepler 框架存在两个问题：①OAI 目前所支持的都是拥有大量对象数据的数据提供者，而 Kepler 框架是针对个人出版者的，每个使用 archivelet 的个人出版者所拥有的对象数据量很小，但是注册的 archivelet 的数量是可以很大的；②由于使用 archivelet 的个人出版者的数据通常存放在个人 PC 上或工作站上，所以不能保证这些机器实时在线。

3. Kepler 架构-Kepler Architecture

这部分将从 Kepler 的架构方面讨论 Kepler 框架的实现，如图 3 所示：

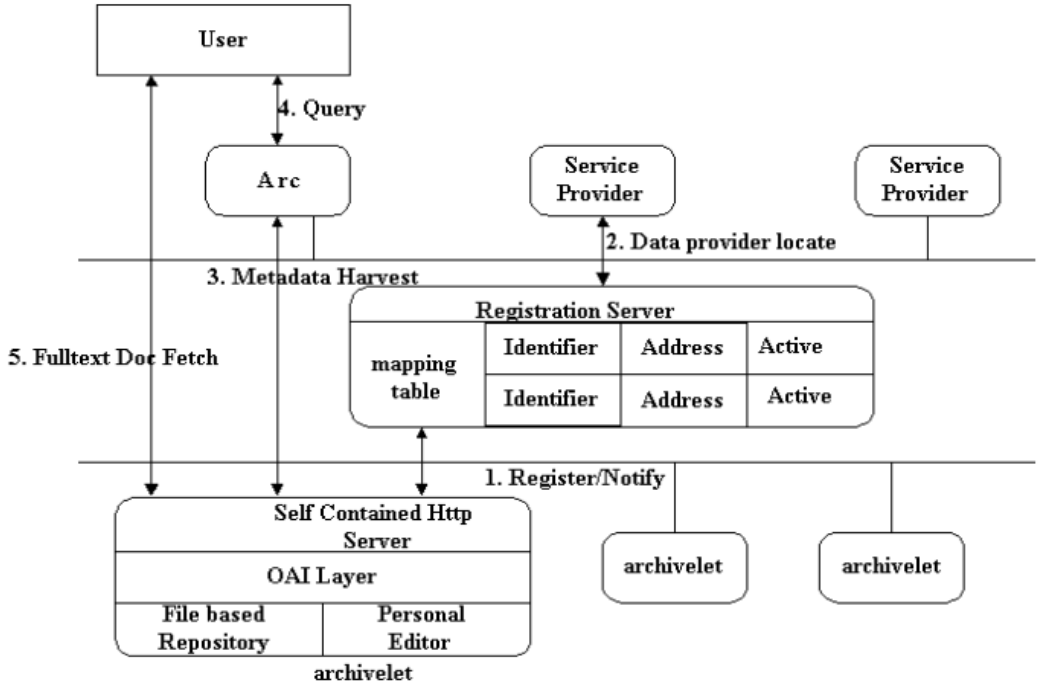

图 3　Kepler 架构

图3中注册服务器允许archivelet向其注册，注册服务器将实时跟踪archivelet的活动/非活动状态信息，也就是说，每个archivelet允许注册服务器知道自己是否在线。注册服务应能处理数千个接入点。服务提供者使用注册服务器定位所有Kepler的archivelet。例如，图3中的Arc是一个发现服务，每天从所有的活动的archivelet收集元数据，这样的服务同时需要知道哪些archivelet是处于活动状态的。在注册服务器中需要维护每个archivelet的标识和它的状态，标识是archivelet当前的IP地址。

Archivelet将OAI兼容的仓储和出版工具捆绑成可下载并和自行安装的组件。采用http传输协议，只支持OAI协议的请求，Kepler的基础服务是发现服务（discovery service）－Arc。包括如下内容：服务和数据提供者之间的一致性、收集计划（harvesting scheduling）、容错处理和数据提供者负载。

当一个读者向Arc发出请求时，不仅需要返回与查询条件匹配的元数据记录，而且也需要获得包含这些元数据记录的archivelet的状态，Kepler针对全文获取采用三种模式：①只提供URL地址；②在archivelet离线之前缓存结果；③缓存经常访问的文档，访问频率是通过所有用户访问Arc和所有注册在Kepler上的archivelet的次数来确定。

4. 原型系统实现

Kepler框架的原型系统的第一步是基于LDAP的注册系统。服务提供者使用的是修改过的Arc。Arc使用Oracle数据对收集到的元数据创建索引。使用OAI协议，服务提供者每天进行收集工作，并更新收集到的元数据。所有注册的archivelet可通过注册服务来维护。出版工具包括简单的文档显示功能和向archivelet上传指定的元数据和文件的功能。出版工具和自动注册处理客户端在一起，这个客户端与服务提供商进行交互，这些工具可以从Kepler网站免费获得。

5. 总结

Kepler的主要目的是为出版和发现特定信息提供灵活的方式。Kepler借鉴了成熟的P2P系统和Web搜索引擎（如Google）方式，并使用标准的服务提供者－Arc为广大的用户提供发现服务。Kepler框架遵循OAI标准，为小型的个人出版者提供了出版分发个人论文资产的简单方式。

参考文献

[1] Kepler Home page. http://kepler.cs.odu.edu:8080/
[2] The Open Archives Initiatives. http://www.openarchives.org
[3] Kepler - An OAI Data/Service Provider for the Individual - Kurt Maly, M. Zubair, Xiaoming Liu,D-Lib Magazine 7(4), April 2001.